Are Kant's categorical imperative and instrumental rationality incompatible? The case for the prisoner's dilemma

Brinca, Pedro

Stockholm University

1 June 2005

Are Kant's Categorical Imperative and Instrumental Rationality Compatible? The Case for the Prisoner's Dilemma*

Pedro Brinca

June, 2005

Abstract

Why is good good and bad bad? Kant's categorical imperative (KCI) and instrumental rationality are analyzed under the game-theoretical framework of the folk theorem. Prescribing different courses of action under the one-shot game, Kant's categorical imperative emerges as instrumentally rational provided that the conditions of the folk theorem are observed and the norms and values underlying KCI are presented as selective advantages of the group of reference in which the individual belongs. Norms and values are argued to be instrumental in nature and KCI and instrumental rationality become two faces of the same coin.

Keywords: Evolutionary game theory, norms, values, prisoner's dilemma, instrumental rationality

JEL: C73, D02, D74, D9

*Bachelor thesis in Economics, Department of Economics, Stockholm University. Advisor: Adam Jacobsson

1. Introduction

"If people are good only because they fear punishment, and hope for reward, then we are a sorry lot indeed."

Albert Einstein in Calaprice (2000)

The purpose of this essay is to find whether Kant's Categorical Imperative (KCI) is compatible with instrumental rationality, and if so, under what circumstances. KCI is often mentioned as a reference on what moral standard is concerned: it classifies courses of action as morally permissible or not. Instrumental rationality is at the heart of Economics and its aim is to model agents' decisions. Its outcome-oriented way of reasoning appears to go directly against KCI. Moral issues were at the birth of Economics. "*An Inquiry into the Nature and Causes of the Wealth of Nations"* Smith (1776), was in the following of Smith's previous book "*The Theory of Moral Sentiments"* Smith (1970), where the author was mainly concerned with the conflict between pursuing self-interest and moral judgments. He derives social behaviour and social institutions from a set of principles he argues are on their origin. This work, laid the psychological foundations for its acclaimed classic. Social norms and institutions are each time more and more used in recent economic research in many different fields, from growth theory to the economics of institutions, from political economics to decision theory.

1.1 Question

Economics' object of study is to "(…) *know how individuals and society decide on what to do with scarce resources, that can have alternative ends, that may be used to produce goods and services and how to share them to be consumed now or in the future* (…)" Andrade (1998). The theoretical framework concerning how agents decide is a corner stone of Economics. For long, in spite of recent advances in behavioural economics, instrumental rationality has been at the foundations of most models being presented. It requires only that the agents order their preferences consistently and act accordingly (maximizing their utility).

Rationality is then presented as an instrument to an end, maximizing welfare. Kant argues that "*The moral worth of an action does not lie in the effect expected from it, nor in any principle of action which requires to borrow its motive from this expected effect*.", Serafini (1989) , This suggests that when deciding, the agent should not take into concern the outcome of its actions. This hardly looks rational in view of instrumental rationality. Is then instrumental rationality, a goal-oriented way of acting, incompatible with KCI? And if it is not, under what circumstances can that be?

1.2 Limitations

Game theory is used as a proxy for instrumental rationality and used throughout this analysis where several simplifying assumptions were made. Symmetric games, strong preferences, common knowledge, perfect information, consistency of preferences and so on. Naturally, the results are expected to hold only within this framework. The analysis focuses mainly on a special game theoretic setting, the Prisoner's Dilemma, which captures the main intuition upon the analysis here being made. Further assumptions are made throughout the paper, most of them in my opinion either reasonable or not crucial for the relevant conclusions drawn.

1.3 Structure

This paper is divided in seven parts. After the introduction, Part two starts by analyzing KCI and the basic axioms from which it is derived. The analysis continues with part three into introducing instrumental rationality and its axiomatic foundations. The fourth part begins with the motivation for the use of game theory as a proxy for instrumental rationality. This part is then divided in another four sections. The first introduces the notation and structure of the classic two person game, defining the concepts used through the other three parts, ending with the presentation of the prisoner's dilemma and its incompatibility with KCI. In the second section, repeated games are introduced and compatibility between KCI and instrumental rationality presented as the cooperative equilibrium, ending with the remark that motivation is in need to consider the Pareto-efficient

outcome, KCI's course of action. The third section explains KCI's foundations from an evolutionary perspective, establishing the interdependency of values, norms and evolutionary success, addressing the problem outlined in the end of the second section. The fourth and last section addresses the implication of the size of communities into cooperation and therefore the compatibility between instrumental behavior and KCI. The fifth part is devoted to some concluding remarks.

2. Kant's Categorical Imperative

"I am never to act otherwise than so that I could also will that my maxim should become universal law. Here, now, it is the simple conformity to law in general, without assuming any particular law applicable to certain actions, that serves the will as its principle, and must so serve it, if duty is not to be a vain delusion and a chimeric vision"

Immanuel Kant in Serafini (1989)

Kant's Categorical Imperative can be found in the framework of Moral Philosophy, a branch of philosophy that, among other issues, addresses the question of what one should do. Kant tries to construct a standard by which all ordinary moral judgements should be measured, in his book "*The Groundwork of the Metaphysics of Morals*". He then proceeds arguing that such a standard must be established *a priori*. To support such a claim, Kant argues that such a principle emerges from the logical relationship between concepts as duty, obligation and good will. He argues that "*Duty is the necessity to act out of reverence for the law*" Serafini (1989). Laws emerge, through a process of public choice, result from common practices, customs, social norms and so on. The concept of duty, even though it is absolute in its own definition, can change from community to community as practices, customs and social norms. Thus, laws also change. Obligations, in this context, are the acts one is required to carry out by moral or legal imperatives as in Zimmerman (1996). The concept of good will entails the

character of autonomous, independent goodness. The will is good by the intent, the motivation, regardless of its consequences. Concepts are themselves *a priori* of observations and therefore also such a standard or principle. As a result of them and the fact that such a principle has to be absolute in nature, it is independent of any *a posteriori* result. The principle he devises is known as Kant's Categorical Imperative (KCI): one should act as one would like it to become a universal law. It is categorical because it applies unconditionally, without references to any end in particular. It is imperative simply because it is a command, a must do instruction. For example, suppose that one agrees with someone about a loan and then defaults on the payment. If when he contracted the obligation, he did not have the intent of keeping it, he is not fulfilling his duty, by failing to conform to his obligation. The maxim, ask for a "loan and then default on the payment " could not be elevated to a common law once it would mean that the one who defaults would also be defaulted in a loan. The essence of immorality is then to act in such a way, that the will reveals the desire that course of action not to be elevated into a common law.

3. Instrumental Rationality

It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love."

Adam Smith (1776)

Instrumental rationality is the building block of most economic theory. It is the outcome of each action that commands what one should do in order to be rational. Rationality is seen in this context as a mean to an end: maximizing utility. It requires that one ranks one's preferences consistently and that one's behaviour should reflect the maximization of one's preferences. The preferences should be logically consistent. The importance of logical consistency was well demonstrated by Bertrand Russell in one of his public lectures. He argued that

from a contradictory set of axioms, everything can be deduced. So, at some point, an individual rose from the audience and defied Russell to prove that from $2+2=5$, that he and the Pope were the same. Russell then argued that if 2+2=5, that means 4=5. So, let's subtract 2 from each side. That gives us 2=3. Transposing, we have 3=2. Now, let's subtract 1 from each side. 2=1. Now, since the Pope and Russell are two different people, and 2=1... The Pope and Russell are one. This contradiction is a good example of why preferences, to be consistent, should follow determined rules, axioms that should be consistent in order to avoid such illogical results. As in Hausman (1992), the model needs the axiom of completeness. Given any two goods, one should be able to compare them in the sense that a relationship of preference can be established. One should be able to tell whether *a* is preferred to *b,* $a \succ b$, *b* is preferred to *a*, $b \succ a$ or finally if the agent is indifferent between *a* and *b*, $a \approx b$. It needs the axiom of transitivity. Given three goods *a*, *b* and *c*, if *a* is preferred to *b* and *b* is preferred to *c*, then, *a* must be preferred to *c*. Assumes also reflexivity of preferences, *a* good is as good as itself, $a \approx a$.

4. Theoretical Analysis

The issue in this analysis is about people making decisions, deciding strategies. In the early days of economic theory, through the works of classics like Adam Smith, the greater good could be achieved through the pursuit of self-interest. The classic reasoning assumed that markets free from intervention would lead to efficiency. Once the equilibrium is achieved, no one could be better off without it being at somebody else's expense. Inefficient markets were seen as a result of some form of market friction, protectionism, government intervention, asymmetric information or unenforceable property rights. What game theory later showed was that things may, in some circumstances, be different. This will be shown within the framework of the Prisoner's Dilemma setup. Inefficient outcomes can arise even in such circumstances. Game theory has since then been

at the core of most analysis. Nowadays it is the main analytical tool towards the analysis of interactions between agents.

4.1 The base line "one shot" prisoner's dilemma case

Having motivated the use of game theory as a relevant tool to analyze strategic interactions between agents, one should define the concepts forward used. These strategic interactions are commonly referred to as games in the literature. As in Nyberg (2003), a game $G(N,S,P)$ is defined by $N=\{1,2,...,n\}$, the set of players (agents), the set of possible actions and the set of players' payoffs. The set of possible actions for each player is the set S_i of strategies s_i available to him. By combining all sets of strategies for all players we get $S=\{S_1 \times S_2 \times ... \times S_n\}$, the strategy space. The payoffs are the outcome of the game, defined for a specific player and a specific strategy profile. The payoff player *i* gets for a strategy combination *s*, is P_{is}. P_s is then the vector listing each player's payoff for a strategy combination *s*. Let's look at the case where there are two players $N=\{1,2\}$, a strategy set for each player *i* with two possible strategy choices $S_i=\{s_{i1},s_{i2}\}$ resulting in a strategy space $S=\{S_1 \times S_2\}=\{(s_{11},s_{21});(s_{11},s_{22});(s_{12},s_{21});(s_{12},s_{22})\}$ of four elements. The payoffs are therefore going to be defined on the strategy space elements: $P(s_{11},s_{21})=(a,b); P(s_{11},s_{22})=(c,d); P(s_{12},s_{21})=(e,f); P(s_{12},s_{22})=(g,h)$. Let's also assume that the agents want to maximize their payoffs, common knowledge, that they decide simultaneously what to do or that they have no information about the other agent's choices and finally that they are fully informed of the structure of the game. Games following these assumptions are referred too as normal form games. They can be more easily interpreted if put in matrix form as in Fig.1:

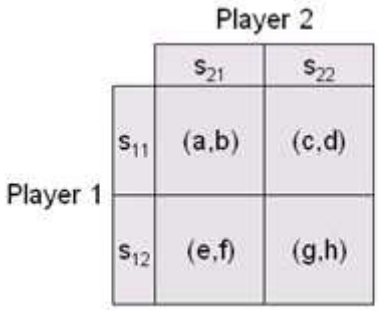

Fig.1

The equilibria are defined as situations where, none of the players can be better of by unilaterally changing their strategy. In such conditions, we refer to them as Nash Equilibria (NE). The outcome of the game is defined by the equilibria strategies and the corresponding payoffs for each player: $\left(s_1^*, s_2^*\right)$and $\left(P_1^*, P_2^*\right)$. For example, $\left(s_{11}, s_{21}\right)$ with the corresponding payoff vector P_s: $\left(P_1\left(s_{11}, s_{21}\right) = a, P_2\left(s_{11}, s_{21}\right) = b\right)$ is a NE as long as $\left(a \succ e\right) \wedge \left(b \succ d\right)$, meaning that the payoff *a* must be preferred to payoff *e* for player one and that the payoff *b* must be preferred to *d* for player two. As referred to before, no player can gain from unilaterally changing their strategies. Games where the payoffs can only be ranked are called ordinal games. In such a case, one payoff can be said to be preferred to another but nothing can be said about the strength of the preference. In the example above one can hypothesize that for player one, the payoff *a* is preferred to payoff *c* but nothing can be said about how many *c*'s player one would have to be provided with to compensate the loss of *a* (marginal rate of substitution). In ranking their preferences, each of the two players, assuming strong preferences only (ruling out the case where one payoff can be equal or better preferred than the other, often referred as being weakly preferred), can do it in $4! = 4 \times 3 \times 2 \times 1 = 24$ ways, all the possible combinations of a chain of hierarchy between preferences. With two players and two possible strategies for each, one would have $4! \times 4! = 24 \times 24 = 576$ different games. It is important then to narrow the scope of analysis. The games that will be under analysis are symmetric, in strategies and in payoffs. The motivation for this comes from the very definition of KCI.

As seen before KCI implies the universality of strategies, in the sense that irrespectively of the payoff, individuals facing the same dilemma should act in the

same way. In the jargon defined above means that players with the same strategy set available to them should make the same strategy choices. For that case, one should then consider the games where the set of available strategies S_i to each player are the same, or formally, that $S_i = \{s_1, s_2\}, \forall i \in N = \{1,2\}$. Another simplification made is that when taking the same strategy choice (in the simpler case with only two strategies) $s = \{s_i\}$, the payoffs are the same i.e.: $P_1(s_i) = P_2(s_i)$ for both players. Let's take a look at the payoff matrix that results from the assumptions made:

		Player 2	
		s_1	s_2
Player 1	s_1	(a,a)	(c,d)
	s_2	(d,c)	(b,b)

Fig.2

This kind of games is said to be symmetric due to the nature of the payoffs' distribution. Under the assumptions above, the number of possible games drastically reduces. Now, as one player ranking the preferences automatically ranks them to the other player, one has only $4! = 4 \times 3 \times 2 \times 1 = 24$ different games.

One type of game will be closely analyzed, the Prisoner's Dilemma. The example is given by two thieves that are caught by the police and are kept in separate rooms. The detectives don't have that many pieces of evidence on them so if none of them confesses, they both get convicted for a minor crime. If they both confess they both get convicted with a reduced sentence (still longer than the one for the minor crime) for having confessed. If one of them confesses and the other does not, the one that confessed walks free and the one that did not confess goes to jail with the longest sentence. We assume that both players want to avoid jail as much as possible (for preference ordering) and that their actions reflect that (agents are rational). They can't communicate and therefore have no information of what the other thieve decided in the meanwhile (what is equivalent to simultaneous decisions) or form binding agreements, and they both are perfectly informed of the structure of the game (common knowledge). In this class of

games, the payoffs' distribution (continuing with the symmetric case) must respect the following chain of preferences: $d \succ a \succ b \succ c$. The outcome $\{P_1(s_1, s_2) = c; P_2(s_1, s_2) = d\}$ is not stable in the sense that player one would increase his payoff by unilaterally changing his strategy from s_1 to s_2 once, by definition of the ranking of preferences, $P_1(s_2, s_2) = b \succ P_1(s_1, s_2) = c$. By symmetry the same applies to the outcome $\{P_1(s_2, s_1) = d; P_2(s_2, s_1) = c\}$ and in this case it is player two who could improve its payoff by unilaterally changing his strategy. Finally, the outcome $\{P_1(s_1, s_1) = a; P_2(s_1, s_1) = a\}$ is not stable either, once both players would gain from changing their strategies. The NE is given by $\{P_1(s_2, s_2) = b; P_2(s_2, s_2) = b\}$, once there is no way of increasing the payoff by any unilateral change of strategy of any of the two players. The special interest of this game is that if they could form a binding agreement, if they could impose a universal law for them to follow (and if it is universal it is by definition the same for everyone facing the same dilemma as it is the case), if they would cooperate, they would promptly agree that their strategy choices would be s_1, once the payoff both would get, *a*, is preferred to the payoff they both get in equilibria *b*. This is a clear case that shows that a NE needs not to be Pareto efficient. At least one person could be made better off without making anyone worse off. In this case, even both would be better off. The desired universal law, in which both players would choose s_1, is not observed and, in this example, KCI and instrumental rationality yield different courses of action, i.e. strategy choices. This particular game is also interesting because the distribution of payoffs reflect much of what takes place in reality. First there are increasing returns to scale in the sense that two players cooperating (choosing the Pareto-efficient outcome) are better off than not cooperating. Then, there is a reward to unilateral defection, what can be seen as some sort of advantage gained or appropriation of part of the other player's investment in the not observed cooperative result. This result, that two individuals following rational courses of action end being worst off is the foundation of the impact game theory had in the dominium of the social sciences, as opposed to the classic invisible hand mechanisms that would conduce markets

to efficiency by themselves. But this analysis only takes into account that players interact once. If time comes into play, the analysis will be different.

4.2 Repeated Interaction

A common approach is that extending the analysis for repeated games, the scope for cooperation, in the sense of following the desired universal law, emerges naturally, purely from instrumental reason without the need of any sort of explicit binding agreement. The problem is that if the preferences are only ordinal, there is no way of taking into account the total utility agents get from repeating a game. This way, the analysis is extended in the sense that the preferences are to be defined cardinally. Now, values (utility measures) are assigned to payoffs. One is not only able to rank the preferences but also to say how much the agent is willing to give up from the outcome of playing strategy 1 for the outcome of strategy 2. Assuming the same game structure as before, a strategy for the repeated game has to be defined. Several could be chosen: one could choose to defect in every period of the game, or cooperate in the first period and then defect in the following periods. A particular strategy, "*Tit for Tat*" has shown to be particular robust, empirically and theoretically, as in Axelrod and Hamilton (1981). In this strategy, agents cooperate on the first round and then simply repeat the other agent's play on the previous interaction. If the game is repeated indefinitely, or has an uncertain end (otherwise, as shown by backward induction, the results won't differ from the one-period game), players will have to sum up the payoffs from period after period. Naturally, future payoffs are expected, for the same amount of utility given in their period, to be less valuable, and that factor should be taken into account when calculating the present value of each of the alternative strategies. From now on, let's call strategy 1 cooperate and strategy 2 defect. Assuming a discount rate of δ per period, a payoff from the n^{th} game will be worth δ^{n-1} times less than the first. From the assumption that future payoffs are less valuable, *ceteris paribus*, it is implied that the value for δ must rely between zero (below zero discount rates would have no economic sense) and one: $\delta \in \Re : \delta \in [0,1]$.

When δ is zero, it means that the individuals do not assign any utility to future payoffs and the game can be reduced to the one interaction case. When δ is one, agents take future payoffs as much into account as current payoffs. As agents are not sure when the game will end, the payoffs are calculated as if the game would be repeated infinitely. Given the assumptions below, player's $i = \{1,2\}$ strategy at interaction (time) *t* is then:

$$\begin{cases} s_{1,t} = Cooperate\ if\ \ s_{2,t-1} = Cooperate \\ s_{1,t} = Defect\ \ otherwise \end{cases}$$

Given this, one can construct the expressions representing the payoff for cooperating in the first round, or defecting from the start, given that the other player starts by cooperating and following the same strategy ("*Tit for Tat"*). As *n* is assumed to be infinite, the result is a geometric progression, converging to the values represented by the expressions below:

$$V_i^C = a + \delta a + \delta^2 a + ... + \delta^n a = a + \frac{\delta}{1-\delta} a$$

$$V_i^D = d + \delta b + \delta^2 b + ... + \delta^n b = d + \frac{\delta}{1-\delta} b$$

The first equation gives the value of cooperating, given that the other player cooperates (the trigger strategy). The second equation yields the value of defecting. As the trigger strategy was to start cooperating, player *i* gets a reward for cheating in the first period but, as seen before, this situation is not stable, once in the next period, the other player will also change its strategy and the payoff will be of *c* for each period (the original static NE), a result of both players defecting from the initial strategy of cooperating. Cooperation is rational as long as the value of cooperating is bigger than the value of defecting:

$$a + \frac{\delta}{1-\delta} a \geq d + \frac{\delta}{1-\delta} b \Leftrightarrow \delta \geq \frac{d-a}{d-b}$$

This last expression $\frac{d-a}{d-b}$ is the threshold value for the discount rate above which cooperating is more rewarding than defecting. The higher the returns to cooperation $(a-b)$, the lower the threshold value for the discount rate has to be in order for cooperation to be achieved, for a given reward *d* for unilateral defection. These kinds of results are commonly known in the literature as Folk

theorems. They show that when time comes into play, at the threat of retaliation, the scope for cooperation increases. In the cases where the agents' subjective discount factor δ is sufficiently big, KCI and instrumental rationality yield the same choice of strategies, therefore being compatible in that sense. An example of increasing scope for cooperation when time comes into play can be devised. Nevertheless, there is literature concerning the fact that under certain assumptions, the scope for cooperation can actually decrease as a function of higher discount rates i.e. if the agents value more future payoffs. Following Skaperdas & Syropoulos (1996) and Skaperdas & Garfinkel (2000), if defecting in the present will lead to an increase in payoffs tomorrow, the more one values these future payoffs, the more prone one will be in considering the benefits of defecting. As the authors point out, a King deciding whether to engage into war or not, has to take into account the strategic implications a loss in the first interaction would imply in the payoffs for the available strategies in the second period and so on. If facing a prisoner's dilemma situation, the King gets cheated by his counterpart and sees the opponent appropriating part of his territory, the payoffs for the same strategies would diminish in the second period, once the resources generated by the appropriated factors are no longer available. Also, the King's strategic strength would diminish in the forthcoming period once it has less resources and the higher the value put upon future payoffs, i.e. the higher the discount rate is, the more scope for conflict. On the other hand, by symmetry, the opponents' decision to cheat in the first period is rewarded with the strengthening of his strategic position on the second period, once, as before, doing better in the first period will make the defecting King to do better in the second period. The rational choice of strategies for both parties will then be to defect right from the first period. Given this game structure (increased payoffs in future games as function of current defection), the time-extended approach will increase the scope for defection.

Back to the traditional folk theorem, once adjusted for a proper discount rate, KCI and instrumental rationality yield the same course of action, suggesting that they are in fact compatible given that the agents value future payoffs above a certain threshold. However, there is a problem with this way of reasoning. Acting

according to KCI also requires that one's choices of acting have a merit of their own. They should not be judged having as reference the utility they would yield. In other words, one's conduct should not be dependent on the outcome. The strategy choice of cooperating satisfies KCI's condition of universality but relies upon the weighting of the payoffs of the possible outcomes. One needs to have a good reason to believe why a Pareto efficient equilibria should be the individual's choice resulting from the desired universal law, other than higher payoffs. A law or norm that, as defined above, would emerge from the axiomatic reasoning of the base concepts of good will, duty and obligation that are in the origin of KCI. These concepts themselves emerge from common practices, customs, social norms and so on, instead of coming directly from weighting utilities and choosing strategies accordingly. If such reason exists, KCI and instrumental rationality can coexist, yielding the same course of action, under the circumstances illustrated above, and show that cooperation may not only be the result of instrumental behavior.

4.3 Social Norms as Selective Advantages

I expect that I have shown how KCI behaviour can, in some circumstances be explained, on what the outcome of the strategies is concerned, as a form of time-extended instrumental rationality. There is room though to clarify why the individuals would, independently from the outcomes, elect a course of action that would increase efficiency (to act in such a way that the equilibrium will be the Pareto-efficient one instead the "prisoner's dilemma trap"). It has been argued, as in Ullmann-Margalit (1977), that the function of moral norms is precisely to enable agents to coordinate and cooperate to achieve the Pareto-efficient results in situations where the pure pursuit of self-interest, in the classic sense, prevents the agents from achieving such an objective. In fact, the importance of attaining efficient results can be seen on how societies have evolved through time. If we look back to the dawn of mankind, it's not hard to realize how cooperation has its merits. Synergy effects (or more technically, increasing returns to scale) made men cooperate in hunting and fighting. A tribe with greater cohesion, greater propensity to cooperation between its members would be a tribe that would have,

ceteris paribus, a greater probability of winning a war. Consequently, any factor that would enforce cohesion, cooperation, would be a selective advantage.

Social norms play an extremely important role in this mechanism. Religion, laws, education and so on, are all mechanisms that enforce cohesion between the members of a society. Even Adam Smith, defending "*invisible hand*" processes, decentralized equilibria achieved by the pure pursuit of self-interest of individuals, safeguarded that some structures need to be present in society for efficiency to be achieved in such a way. He stressed the importance of property rights and general observance of social norms such as the interdiction of stealing or misrepresentation. The Ten Commandments are a good example of a code of conduct that obeys to KCI reasoning.

Agents in the prisoner's dilemma case would even be willing to give some of their Pareto-efficient payoffs in order to ensure a better outcome. In the picture below, the grey area represents all the possible payoffs where any of the agents would be better off at no expense to the other player.

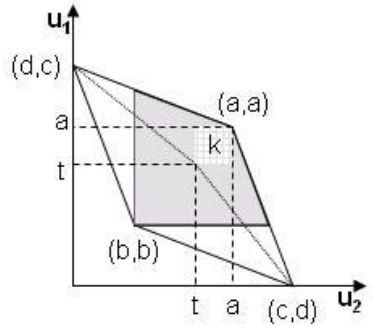

Fig.3

In this example, it is considered that the payoffs are infinitely divisible, a simplification that is necessary for the picture above to be consistent. Keeping symmetry of payoffs for simplification purposes, any combination of payoffs belonging to the segment $\overline{(b,b)(a,a)}$ would be preferred to the outcome (b,b). For example, both agents would be willing to give up $a-t$ of their individual Pareto-efficient payoff in order to pay for some kind of mechanism that would enforce cooperation. The area $k=(a-t)^2$ represents then the demand for enforceable, credible, binding regulations (in the form of institutions, norms, social pressure) for a given alternate possible payoff of t for both players. Not only are these

factors selective advantages, but also a rational inherent desire from the agents. As in Thoreau (1849), "*Let every man make known what kind of government would command his respect, and that will be one step toward obtaining it.*" It's only natural that throughout the ages they are so thoroughly institutionalized. As they are internalized into their personal beliefs, generation after generation, it is more likely that the agents' values and concepts as the ones defined above (good will, duty and obligation) will reflect cooperative behaviour. Even though the channels of intergenerational transmission of values and norms may be innumerable, it is reasonable to assume that many are acquired by observation, by reverence towards references' (family, teachers, preachers and sport stars or rock singers and so on..) behavior, by what is part of the status quo. At this point, values and norms will shape agents' wills in the Kantian sense, and will therefore elect courses of action, strategies that are trendily efficient. Institutions are designed to ensure the stability of such state of things. Today virtually any society has courts, police, schools, churches, all of which perfect examples of mechanisms that are designed to ensure that, in a compelling way, individuals are integrated socially, that they behave in a KCI perspective.

The fact that given strategies guided by instrumentally driven factors (inherent instrumental rationality or evolutionarily selected efficiency enhancing values and norms), are selective advantages is a common approach in evolutionary game theory. In particular, certain strategies may reveal themselves to be more successful than others. In the late 1970's, Robert Axelrod held a tournament where intellectuals from a wide range of fields participated in repeated prisoner's dilemma games of approximately 200 rounds. Each contestant had its proposed strategy played against the other proposed strategies in five period games. The strategy that had the highest score was submitted by Anatol Rapoport and it was precisely the "*Tit for Tat*" strategy. A good way to understand how this is meaningful in the evolutionary systems context is to observe the Spatialized Prisoner's Dilemma, as in Grimm (1997). In this case, the space is represented by a surface of contingent squares. To each square is randomly assigned one of eight different types of players. Each player type has a distinctive strategy profile. Games with the structure mentioned above (repeated

games with uncertain ends and prisoner's dilemma structure of preferences) are played between each player and all his neighbours (players in contingent squares). In the end, scores are tallied and if a player gets a score that is lower than any of the contingent players, it replicates the strategy of the highest scoring player in the next round. If no contingent player had higher scores in that interaction, the strategy profile is kept and applied again in the next interaction. In the following example eight types of agents, each one with its own different strategy, represented by eight different colours compete for the dominance of the squares. Screenshots of the playing surface from Grimm (1997) show that in an early stage of the competition, strategies that privilege defection, (shown in white and light grey), tend to do better in the first rounds, meaning that there are more agents playing those strategies, as it can be seen in the first two screenshots:

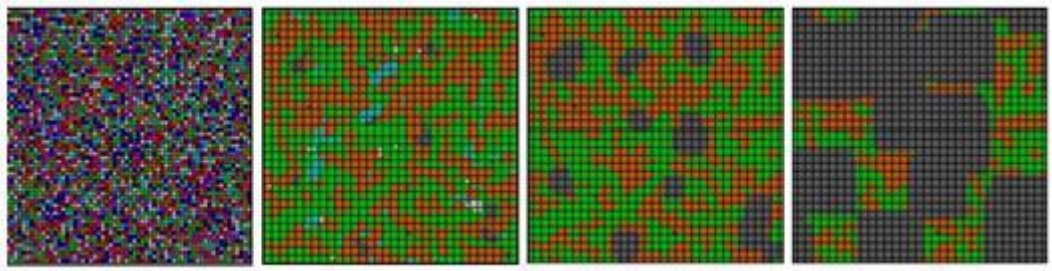

Fig. 4

However, after some rounds, "*Tit for Tat*" (dark grey) starts to prevail as it shows in the 3rd and 4th screenshot, and after twenty six interactions, all the plane has been "conquered" by the agent playing "*Tit for Tat*". This result is however, sensible to the initial conditions. Given the structure of this "experiment", it cannot be proven that "*Tit for Tat*" will always be the dominant outcome. In fact other strategies can proliferate and dominate the plane. But "*Tit for Tat*"" has proven to be particularly robust. By the same way, some values, rules of conduct can lead to reproductive success among the populations that share them or to oblivion. Populations with values that defend private property and entrepreneurship may lead to capital accumulation and development of knowledge and consequently to technological superiority against populations that loath those same concepts. What the Spatialized Prisoner's Dilemma example suggests is that cooperation is in fact superior in the long run. But "Tit-for-Tat" is not a naïve form of cooperation. It starts by cooperating but retaliates (by defecting,

something that lowers the opponent's payoff) if the agents with whom it interacts defect. An analogy can be made with everyday life. In reality, people tend to follow the norm, at the threat of retaliation from society if going against the norm, through the same norms and institutions (referred above) , that as seen before are not only selective advantages but are also instrumentally rational desires from the agents. In this framework, KCI and instrumental rationality are compatible, as long as agents value future payoffs at an equal or greater rate than δ. But can anything be said about the actual value of δ? If agents are in general overly impatient, in the sense that their subjective discount factor is below the modelled threshold, cooperation will never be observed even in repeated games and there is no scope for compatibility between KCI and instrumental rationality. Let's assume then that the utility each strategy choice yields can be affected by these norms and institutions without altering the structure of the game i.e. the ranking of preferences. Remember that a prisoner's dilemma problem requires only that preferences are ranked in an ordinal way. If the norms and institutions devised impute penalties on cheating, the value of defecting will be lower, and therefore lower will be the real δ. Overall efficiency would be easier achieved and again, natural selection would ensure them to last. If institutions and norms would not reflect such an effect they would be pointless, in this framework. Their existence in all societies is the best proof of their importance.

There is an alternative way through which the above mentioned factors affect the outcome of agents' interactions. One can assume they alter the structure of preferences and therefore alter the structure of the game. Let's try to picture the following example: Pedro has bread at his place and Adam has water. They have the option of either trade (T) some bread for some water and they both get a portion of both resources, (bw, bw), or steal (S) the other one's resource. If Pedro intends to trade and gets robbed, he will have the worst outcome possible, being thirsty and hungry and Adam will have the best outcome, having all bread and water to himself, $(0, BW)$. The symmetric result occurs if Pedro is the thief and Adam the one that decided to trade $(BW, 0)$.If they both rob each other, they will

have each other's resource, either being hungry or thirsty, (b, w). The preferences' ranking is the same for both:

$$All\ Bread\ \&\ All\ Water \succ Some Water\ \&\ Some\ Bread \succ Only\ bread\ or Water \succ None$$

This is a classic Prisoner's dilemma. As before, one can devise a payoff matrix to better illustrate which will be the equilibria:

		Pedro	
		T	S
Adam	T	(bw,bw)	(0,BW)
	S	(BW,0)	(b,w)

Fig.4

By instrumental rationality, both individuals will try to rob each other and fall into the Prisoner's Dilemma trap, ending both with each other's resource. If they would cooperate with each other, by trading, they both would be better off. Now, suppose they both go to church on Sunday, and the word spreads around. Stealing is seen as a sin. The moral values of their community, which condemn stealing, will make stealing to be less rewarding, because Pedro and/or Adam will suffer from social discrimination, guilt complex or even jail possibility. That disutility can be reflected in the payoffs distribution. Now, the action of stealing will result in a specific decrease of utility to the outcome from that strategy. If the penalty is sufficiently high to alter the structure of preferences, a new game will arise. If it were not sufficient to alter the structure of preferences over a large enough number of people, and assuming that social norms also come from tradition and customs, the norm would become so widely violated that would, ceteris paribus, fade away. Suppose that the loss in utility implied in stealing (affects all payoffs resulting from the use of that strategy), makes the result from having some water and some bread (that can only result from trade) preferable to having all the water and bread (that can only be achieved by unilateral theft). The preferences will now become:

$$Some Water\ \&\ Some\ Bread \succ All\ Bread\ \&\ All\ Water \succ Only\ bread\ or Water \succ None$$

Now there are two NE, where they both trade and where they both steal. Consequently, the penalty imposed is not enough to enforce cooperation even though it changed the structure of preferences. Which equilibria will then prevail? Through a static approach and pure strategies, nothing can be said. But repeated games applied to this case, quantifying utilities (cardinal games) and again from the assumption they both play a trigger strategy where they are expected to trade in the first game and then do whatever the other one did in the previous game ("*Tit for Tat"*), will unambiguously result in cooperation once that for any value of the payoffs, as long as trade is marginally preferable to the unilateral defection reward, cooperation is always preferable to defection. Through this reasoning, and admitting repeated games are more plausible to observe in reality rather than static games, instrumental rationality and KCI behaviour would induce the same choice of strategies, regardless of how agents subjectively value future payoffs. Again, social norms, values and institutions are presented as selective advantages increasing efficiency. Even though they imply only loss of utility in the outcomes, the change in the strategic context makes the overall result more efficient. With greater resources available to them, groups would be more successful in terms of natural selection and transmit those cultural factors to the forthcoming generations. Even though that seems to be the general case, exceptions can be thought of. There may still be values and norms that in some situations may apparently induce inefficient outcomes. In the following example, a patient with severe brain damage and in a permanent coma, left expressed in her will that if found in such conditions, would like to have shut down any artificial mean of life support. The patient revealed her preference and society would save resources if the life support mechanisms would be shut down. Nonetheless, life support could be maintained for a long period, due to values as the inviolability of human life, and the obligation to fulfil the duty of assistance.

4.4 The importance of the size of communities

From the reasoning above one would be eager to assume that the natural outcome for this entire story would be to have one homogenised society, with one set of norms or values, reasoning in the same way. As in Alesina & Spolaore

(1997) there are several benefits of larger societies or countries. As population size increases, per-capita cost of public goods decreases and there are increases in productivity from benefiting from larger markets. The increasing returns to scale assumption made above also helps. If we look to the prisoner's dilemma case, and instead of assuming two players interacting, we assume two group leaders, deciding on whether to cooperate or not in commerce for example, the same reasoning above can be applied to defend that groups will cooperate. By cooperating, they tend to build common laws, practises, customs and institutions. Most theories in the framework of international economics, stress the overall gains from cooperation (trade) between economic blocks, and again, institutions were created and policies implemented to walk in such direction. Historically, this increasing homogeneity cannot be denied. The increasing network of commercial and cultural relationships between all countries can rely also in the progressive strengthening of international law. Organizations like the United Nations and the World Trade Organization might seem powerless in same cases in the present day, but from a historical perspective their influence is not comparable with anything seen before in terms of enforcing international common rules upon the world, either on the credibility of the threat of enforcement and the extent they can reach.

Even though it seems to trend in this way, there are factors that may slow down the process or even halt it. Alesina and Spolaore (1997) argue that the greater the number of members of a population, the harder it is to be homogeneous. Even considering that there are increasing returns to cooperation, there will be a number of individuals within a population from which they will have to occupy a so vast area (using a particular notion of distance to be better explained below), that interaction between all individuals will become more difficult and heterogeneity will grow. As Aristotle once said, "*experience has shown that it is difficult, if not impossible, for a populous state to be run by good laws*" Saunders and Sinclair (1981). Heterogeneity, or diversity of preferences, can lead to serious economic and political setbacks. Easterly and Levine (1997) show how ethnic diversity interferes with good governance. To better illustrate the impact that the size of a community can have in cooperation, let's refer to the repeated game example given above. As it was demonstrated, the scope for

cooperation increases with the time-extended approach. But the model is limited in the sense that players are expected to play the game *ad eternum* which is not very reasonable, or that they are uncertain about when the game ends. As nothing is said about the effect of the number of individuals in a group upon the probability of the interactions being more frequent, the model is extended to incorporate such an effect. Now, future payoffs are discounted not only by the subjective discount rate (how agents "value" time) but also by the probability agents assign for them to interact more than once. Again, symmetry is assumed upon those expectations, implied by the common knowledge assumption. Let's assume that there is a probability p that the game will end at the second interaction. This can be interpreted like the probability of meeting two times with player two and repeat the interaction. For the payoff of the first interaction, p is one, once the player is actually facing the interaction. Then, the probability of the agents interacting is increasingly smaller, in the sense that it is more probable that the agents interact twice than three times and so on. For simplification purposes it is assumed that the probability value assigned for each period decreases in the same functional form of the discount rate, once it respects the condition of monotonicity.

Now, the expected values for cooperation and defection are:

$$E\left[V_i^C\right] = a + \delta.p.a + p^2\delta^2 a + ... p^n\delta^n a = a + \frac{p\delta}{1-p\delta}a$$

$$E\left[V_i^D\right] = d + \delta.p.b + \delta^2 p^2 b + ... + \delta^n p^n b = d + \frac{p\delta}{1-p\delta}b$$

The threshold value for the discount rate, which promotes cooperation, is then given by:

$$a + \frac{p\delta}{1-p\delta}a \geq d + \frac{p\delta}{1-p\delta}b \Leftrightarrow \delta \geq \frac{d-a}{p(d-b)}$$

As by definition of probability, $p \in [0,1]$ even though in this case, one has to exclude the possibility of p being zero for the expression to be meaningful and of one, else wise one would be back in the traditional folk theorem case. When p is zero, it means that we are back at the initial static game, once that the probability of interacting more than one time is zero. For any other value of p, the denominator is smaller and therefore, the value for the discount rate increases.

Agents have to give higher value to the future in order to cooperate. In other words cooperation is harder. But, an increase in the discount rate cannot offset a decrease in the probability of the agents meeting again *ad eternum*. If $p = \frac{d-a}{d-b}$, the discount rate would have to rise up to $\delta = \frac{d-a}{\frac{(d-a)}{(d-b)}(d-b)} = 1$, meaning that future payoffs would have to be as much valuable as current payoffs, which does not make sense. If the probability is even lower, future payoffs would have to be more highly regarded than current payoffs. It is only reasonable to assume that below the threshold value for *p*, cooperation is no longer viable, no matter how highly agents value their future payoffs.

Thus, to make the connection between the probability of the interaction and *N*, the number of people inside a certain group, the probability is now going to be a decreasing function of the number of individuals that constitute the reference group: $p(N) = \frac{1}{N}, N \geq 2$, meaning that the higher the number of agents within a group, the lesser the probability the same two agents will meet again.

As before the value for each strategy is given by:

$$V_i^C = a + \delta.\frac{1}{N}.a + \left(\frac{1}{N}\right)^2 \delta^2 a + ... + \left(\frac{1}{N}\right)^n \delta^n a = a + \frac{\frac{1}{N}\delta}{1-\frac{1}{N}\delta}a$$

$$V_i^D = d + \delta\frac{1}{N}b + \left(\frac{1}{N}\right)^2 \delta^2 b + ... + \left(\frac{1}{N}\right)^n \delta^n b = d + \frac{\frac{1}{N}\delta}{1-\frac{1}{N}\delta}b$$

Cooperation is viable when:

$$a + \frac{\frac{1}{N}\delta}{1-\frac{1}{N}\delta}a \geq d + \frac{\frac{1}{N}\delta}{1-\frac{1}{N}\delta}b \Leftrightarrow \delta \geq \frac{d-a}{\frac{1}{N}(d-b)} \geq N\frac{d-a}{d-b}$$

For a unit increase in population $(\Delta N = 1)$, the threshold value for δ increases by $\frac{d-a}{d-b}$, and following the same reasoning as before, enforcing

cooperation becomes harder and harder as the population grows. Cooperation is no longer feasible once the population number is equal or greater than $\frac{d-b}{d-a}$, the value for N above which the discount rate becomes greater than 1. This result suggests that cooperation in a random interaction is more probable in smaller communities (smaller N) than in big communities (larger N) for a given discount rate, what seems in-line with reality. This approach presents a fixed population threshold for a given value of the discount rate, above which no cooperation would be observed. However, the critical value for the discount rate as a function of the size of population depends on the returns to cooperation, $(a-b)$ given that the returns to unilateral defection remain unchanged. As these increase, the threshold value for the discount rate $N\frac{d-a}{d-b}$, below which no cooperation is viable, decreases. If the condition for cooperation (agents valuation of future payoffs equal or greater than δ) was satisfied before, the size of the community N can increase, at least until $N\frac{d-a}{d-b}$ equals the previous threshold discount rate δ without jeopardizing cooperation. For example, with the Industrial Revolution, productivity increased a great deal. It seems reasonable to assume, that the returns to cooperative behaviour increased also with such productivity increases, what would then lead to the sustainability of greater populations.

An alternative way of allowing technological progress to have an impact in the sustainability of cooperation in the model is to assume that the probability of a repeated interaction p is not only a function of the population size but also of technology progress, $p(N,A)$. Assuming that the size of the population affects negatively the probability of repeated interaction, $\frac{\partial p}{\partial N}<0$ and that the technological progress affects it positively, $\frac{\partial p}{\partial A}>0$ the sustained population size for a given discount rate depends on the relative effects of these variables. This approach has the merit of not relying in specific functional forms, maintaining the same reasoning. Nevertheless, in both cases, the case for compatibility with KCI decreases with the introduction of probabilities. Furthermore, it is reasonable to

admit that the increasing returns to cooperation cannot hold *ad infinitum*. There are significant agency costs, and it is only natural to expect that, after a certain point, the returns to cooperation might be, at least in part, offset by those. All this suggests that there is a limit, also dependent on technological factors, beyond which no cooperation would be expected. This comes from the smaller probability of repeated interaction in greater communities. With less repeated interaction and cooperation, diversity is likely to grow once, as said before, it is isolation that creates diversity. The motivation for the probability of interaction being positively related to technology progress comes from the relation between distance and isolation. Distance in this context, means how far a representative individual's actions can reach during his lifetime. According to this concept, the revolution on transportation and communication made distances shorter and dramatically minimized the effect of this variable during the last 500 years. If we go back 20000 years ago, the cavemen or nomads could only spread their influence by the distance they could cover on foot. In one day, he could at maximum cover say 50km. 1500 years ago, men horseback riding, could cover 300km in one day. After the industrial revolution and the railroad, 15000km could be daily achieved. Today, in one day, one can go around the world. This largely diminishes the effect of distance upon isolation and therefore the "creation" of diversity. The assumption that technological progress affects positively the probability of interaction as has a positive effect in cooperation holds as long as it is assumed that it affects agents in the same way, i.e. that symmetry is reasonably kept. A technological breakthrough exclusive only to one agent or one community may alter the strategic characteristics of the interactions and thus have the opposite effect, making defection more attractive. The threat of retaliation that is at the base of the evolutionary success of cooperation may become obsolete due to an unbalance in the strategic strength between the communities. Vikings used the long boat to reach farther lands and pillage other populations. But they did so, precisely because of the asymmetry of technological development on what the knowledge of the sea and crafts were concerned. The threat of retaliation was kept to a minimum because the pillaged populations had few if any possibility at all of

chasing the Vikings. In this strategic context, with no serious threat of retaliation at sight, technological progress might have precisely the opposite effect.

Technology improves the ability to spread influence in another important way: information technology. As writing was invented, knowledge could pass not only between individuals but also from generation to generation in a more accurate and consistent way. Nonetheless, it was only with the printing press that a true revolution came in the massification of coherent information. In the dusk of the XIX century, as the first radio signals were transmitted, information started to travel at the speed of light. The 30's brought us television and advertising. The 80's the internet. Today, in this global village, South Koreans eat at McDonald's and South Africans read The Economist. Again, it is isolation that creates diversity. Finally assuming that populations with different social norms tend to be less interdependent and that may even fight for dominance of their values, again the size of communities has a central role. In the beginning, greater cohesion will make a population more successful towards others. This will increase the number of individuals that constitutes it, either by reproductive success, either by assimilating other populations into theirs. However, as the society grows and becomes stronger, it occupies a larger area (using the distance notion referred above) and diminishes the probability of repeated interaction between the same agents. This heterogeneity can be the source of much political unrest as agency costs get higher and control more difficult. This reasoning raises the idea that countries are more heterogeneous outside (when compared to other countries) than they are inside (when analyzed from its own regions perspective), on what geography and social norms is concerned what is obviously in line with reality. Social norms that embrace values promoting efficiency and the selective advantage that they bring upon greater cohesion, make at least to some degree, KCI's behaviour to predominate, precisely because the concepts that are on its origin, are a result of instrumentally driven mechanisms.

5. Concluding remarks

The aim of this paper is to assess the scope for compatibility between KCI and instrumental rationality. They are opposite in nature. KCI conduct relies on

the merit of the action disregarding whatever utility that particular course of action might give. Reason and autonomy lead the individual to achieve moral righteousness. On the other hand, in instrumental rationality, people's behavior is supposed to reflect the act of maximizing the utility the taking of those actions yields. Reason exists in this framework as an instrument to achieve utility. From this perspective, there is little room for compatibility between these two concepts. People tend to be instrumental in nature and rule their everyday lives by choosing outcomes taking into account the utility they yield. Nevertheless, moral also plays a part in people's decisions. People share values and their preferences are influenced by them. Norms and institutions are supposed to be the materialization of those same values. Once established a link between the Kantian merit of an action and the payoffs such actions yield, the possibility of compatibility arises. The basic axioms where KCI is derived from are a function of people's customs, traditions and so on. These customs and traditions are presented as determinant in the evolutionary success of a population. Their existence implies the establishment of links of cooperation between people. Some values can be "better", in evolutionary terms, than others. Where some values and norms can lead to reproductive success of populations, others can lead to oblivion. The Spatialized Prisoner's dilemma is a good illustration of such mechanisms. Once established that the long run evolutionary success of a population is affected by such values and norms, it is only natural to expect individuals to reflect more efficient forms of behaviour. The link between values and norms and efficiency conducive behavior is the key to assure that KCI and instrumental rationality can be compared in the first place. One of the interesting features of the classic Prisoner's Dilemma is that, following the reasoning above, KCI and instrumental rationality invite to different courses of action. They are clearly incompatible, but when time comes into to play, cooperation becomes more likely and the scope for compatibility increases. The introduction of the population size dependent probabilities in the repeated game analysis increases the realism of the model and shows that the probability of cooperative behavior in a random interaction is higher in a small size community than in a bigger one, something that seems in-line with reality.

6. Acknowledgments

I would like to thank to Adam Jacobsson for his infinite patience, determinant guidance and constant availability; Michael Lundholm for the suggestion of the research question and determinant words of guidance; Mårten Larsson for all the intensive help and support and finally the Department of Economics as whole, for the magnificent opportunity of studying at Stockholm University. Dedicated to Ana Isabel.

7. References

Alesina, Alberto and Enrico Spolaore (1997), "*On the Number and Size of Nations*", Quarterly Journal of Economics, Vol.107,, nº4 1027-1056.

Andrade, João Sousa (1998) Introdução à Economia, Coimbra: Minerva, pp.I-8.

Einstein in Calaprice, Alice (2000), The New Quotable Einstein, Princeton University Press.

Easterly W. and R. Levine (1997), *Africa's Tragedy*, Quarterly Journal of Economics, 112, November.

Grim, Patrick, (1997) *The undecidability of the spatialized prisoner's dilemma*, Theory and Decision, January, vol. 42, no. 1, pp. 53-80(28)

Grim, Patrick, *Undecidability in the Spatialized Prisoner's Dilemma: Some Philosophical Implications, available at:*
http://www.sunysb.edu/philosophy/faculty/pgrim/SPATIALP.HTM#S1

Hausman, Daniel (1992). The Inexact and Separate Science of Economics, Cambridge: Cambridge University Press, pp. 52-54

Kant, Immanuel, *Good Will, Duty, and the Categorical Imperative*, ed. Anthony Serafini, Ethics and Social Concern (New York: Paragon House Publishers, 1989), p. 29.

Nyberg, Sten (2003), Lecture notes on Industrial Organization, Stockholm University, Department of Economics.

Skaperdas, Stergios and Constantinos Syropoulos (1996), *Can the shadow of the future harm cooperation?* Journal of Economic Behavior and Organization, vol. 29 pp.355-372.

Skaperdas, Stergios and Michelle Garfinkel (2000), *Conflict Without Misperceptions or Incomplete Information: How the Future Matters*, University of California-Irvine.

Smith, Adam (1776/1965), An Inquiry into the Nature and Causes of the Wealth of Nations, New York: Modern Library,Saunders, Trevor and T.C. Sinclair (1981), The Politics, Penguin Classics.

Ullmann-Margalit, Edna (1977). The Emergence of Norms. New York: Oxford University Press.

Smith, Adam (1790). The Theory of Moral Sentiments, London: A.Millar, Sixth Edition.

Thoreau, Henry (1993). Civil Disobedience and other Essays. Dover Thrift Editions.

Zimmerman, Michael J., (1996), *The* Concept of Moral Obligation, New York: Cambridge University Press.